证　　明

本证明之附件是向本局提交的下列专利申请副本

申　请　日：　2001 08 03

申　请　号：　01 1 23845.3

申请类别：　发明

发明创造名称：　格式文档中的信息的抽取装置及抽取方法

申　请　人：　富士通株式会社

发明人或设计人：黄晓宏；徐国伟

中华人民共和国
国家知识产权局局长　王景川

2004 年 2 月 17 日

6

权 利 要 求 书

1. 格式文档中的信息的抽取装置，包括：输入格式文档的输入单元（1）；对输入的格式文档进行分析，并保持特殊排印信息的排印信息保持单元（2）；对于分析的结果，利用字号、字体、颜色等排印信息来识别特殊字符串的特殊字符串判定单元（3）；抽取识别出来的特殊字符串的特殊字符串抽取单元（4）；以及输出抽取出来的字符串的输出单元（5）。

2. 权利要求 1 所述的格式文档中的信息的抽取装置，其特征在于，上述特殊字符串判定单元（3）利用格式文档的排印信息，当判断出某个字符串的排印信息为特殊排印信息时，将其判断为特殊字符串。

3. 权利要求 1 或 2 所述的格式文档中的信息的抽取装置，其特征在于，上述格式文档为 HTML 文档，上述特殊字符串判定单元（3）根据对 HTML 文档的分析结果，当判断出某个字符串的字号与周围相比为最大时将该字符串判断为特殊字符串。

4. 权利要求 1 或 2 所述的格式文档中的信息的抽取装置，其特征在于，上述格式文档为 HTML 文档，上述特殊字符串判定单元（3）根据对 HTML 文档的分析结果，当判断出某个字符串的颜色和字体与周围相比为特殊时将该字符串判断为特殊字符串。

5. 权利要求 1 或 2 所述的格式文档中的信息的抽取装置，其特征在于，上述格式文档为 HTML 文档，上述特殊字符串判定单元（3）根据对 HTML 文档的分析结果，当判断出某个字符串的字体与其他不同且为粗字，与周围相比为特殊时将该字符串判断为特殊字符串。

6. 权利要求 1 或 2 所述的格式文档中的信息的抽取装置，其特征在于，上述格式文档为 HTML 文档，上述特殊字符串判定单元（3）根据对

HTML 文档的分析结果，当判断出某个字符串的颜色与其他不同且为粗字，与周围相比为特殊时将该字符串判断为特殊字符串。

7. 格式文档中的信息的抽取方法，包括以下步骤：输入格式文档的步骤；对输入的格式文档进行分析，并保持特殊排印信息的步骤；对于分析的结果，利用字号、字体、颜色等排印信息来识别特殊字符串的步骤；抽取识别出来的特殊字符串的步骤；以及输出抽取出来的字符串的步骤。

8. 权利要求 7 所述的格式文档中的信息的抽取方法，其特征在于，在上述识别特殊字符串的步骤中利用格式文档的排印信息，当判断出某个字符串的排印信息为特殊排印信息时，将其判断为特殊字符串。

9. 权利要求 7 或 8 所述的格式文档中的信息的抽取方法，其特征在于，上述格式文档为 HTML 文档，在上述识别特殊字符串的步骤中根据对 HTML 文档的分析结果，当判断出某个字符串的字号与周围相比为最大时将该字符串判断为特殊字符串。

10. 权利要求 7 或 8 所述的格式文档中的信息的抽取方法，其特征在于，上述格式文档为 HTML 文档，在上述识别特殊字符串的步骤中根据对 HTML 文档的分析结果，当判断出某个字符串的颜色和字体与周围相比为特殊时将该字符串判断为特殊字符串。

11. 权利要求 7 或 8 所述的格式文档中的信息的抽取方法，其特征在于，上述格式文档为 HTML 文档，在上述识别特殊字符串的步骤中根据对 HTML 文档的分析结果，当判断出某个字符串的字体与其他不同且为粗字，与周围相比为特殊时将该字符串判断为特殊字符串。

12. 权利要求 7 或 8 所述的格式文档中的信息的抽取方法，其特征在于，上述格式文档为 HTML 文档，根据对 HTML 文档的分析结果，当判断出某个字符串的颜色与其他不同且为粗字，与周围相比为特殊时将该字符串判断为特殊字符串。

说　　明　　书

格式文档中的信息的抽取装置及抽取方法

技术领域

本发明涉及从输入的文档，例如进行网上销售的网页中自动地抽取出特殊字符串的文档中的信息的抽取装置及抽取方法。

背景技术

现有的从文档中抽取信息的装置，例如有 S. Soderland "Learning to Extract Text-based Information from the World Wide Web" , Proc. 3rd Intl Conf. on Knowledge Discovery and Data Mining (KDD-97)中公开的技术。在现有技术中，利用位于特殊字符串之前的属性名（例如"商品名"）的字符串来判别特殊字符串并将其抽出。

在现有技术中，因为是利用位于特殊字符串之前的属性名（"商品名"等）的字符串来判别特殊字符串并将其抽出的，因而在像'商品名：モノグラムアクセサリーポーチ'那样的、齐备了作为属性名的'商品名'和作为属性值的商品名称的场合是有效的。但是，像因特网的网页那样的文档有各种各样的格式，存在着没有属性名的情况。例如，存在着只有'モノグラムアクセサリーポーチ'的情况。在没有属性名的情况下，采用上述技术就不能抽出特殊字符串。另外，在现有技术中需要人工提供样本供机器学习，不能自动地抽取出特殊字符串。

本发明是为了解决上述问题而作出的，其目的在于提供一种能够从输入的格式文档中自动地抽取出特殊字符串的文档中的信息的抽取装置及抽取方法。

发明内容

为了解决上述问题，本发明的格式文档中的信息的抽取装置，包括：输入格式文档的输入单元；对输入的格式文档进行分析，并保持特殊排印信息的排印信息保持单元；对于分析的结果，利用字号、字体、颜色等排印信息来识别特殊字符串的特殊字符串判定单元；抽取识别出来的特殊字符串的特殊字符串抽取单元；以及输出抽取出来的字符串的输出单元。

本发明的格式文档中的信息的抽取方法，包括以下步骤：输入格式文档的步骤；对输入的格式文档进行分析，并保持特殊排印信息的步骤；对于分析的结果，利用字号、字体、颜色等排印信息来识别特殊字符串的步骤；抽取识别出来的特殊字符串的步骤；以及输出抽取出来的字符串的步骤。

若采用本发明，因为对输入的格式文档进行分析，利用字号、字体、颜色等排印信息来判断出特殊字符串信息并抽取特殊字符串，故能够从输入的格式文档中自动地抽取出特殊字符串，并能够大幅度提高抽取准确度。另外，在现有技术中需要人工提供样本供机器学习，而本发明不需要学习样本，能够对于不同类型的格式文档自动地进行判断和抽取。

附图说明

图 1 为本发明的格式文档中的信息的抽取装置的结构框图。

图 2 为说明本发明的实施例 1 的文档数据和流程图。

图 3 为说明本发明的实施例 2 的文档数据和流程图。

图 4 为说明本发明的实施例 3 的文档数据和流程图。

图 5 为说明本发明的实施例 4 的文档数据和流程图。

具体实施方式

图 1 为本发明的格式文档中的信息的抽取装置的结构框图。

在图 1 的格式文档中的信息的抽取装置中，1 为输入格式文档的输入单元；2 为利用某种方法对输入的格式文档进行分析，并保持特殊排印信息的排印信息保持单元；3 为对于分析的结果，利用字号、字体、颜色等排印信息来识别特殊字符串的特殊字符串判定单元；4 为抽取识别出来的特殊字符串的特殊字符串抽取单元；5 为输出抽取出来的字符串的输出单元。

下面，参照图 2-图 5，以从 HTML（超文本标志语言）文档中抽取出特殊字符串为例来说明本发明的格式文档中的信息的抽取装置的动作。

（实施例 1）

图 2 为说明本发明的实施例 1 的文档数据和流程图。其中，图 2(a)为某个网上销售信息（HTML 形式的文档）；图 2 (b)为图 2(a)中的信息的 HTML 源文件；图 2(c)为实施例 1 的信息抽取动作的流程图。

下面说明实施例 1 的信息抽取动作的流程。在步骤 101 中，输入图 2 (b)所示的 HTML 源文件。在步骤 102 中对在步骤 101 中输入的 HTML 源文件进行分析，发现排印信息。接着在步骤 103-107 中进行特殊字符串的抽取。

首先，在步骤 103 中根据步骤 102 的分析结果确定字符串判断对象。在步骤 104 中判断在步骤 103 中确定的字符串的字号与周围相比是否为最大。若判断为否则进入步骤 106。在步骤 106 判断该字符串的排印信息是否超出了预先设定的范围，如果超出了预先设定的范围则进到步骤 107，结束信息抽取动作。在步骤 106 中如果判断为没有超出预先设定的范围则返回步骤 103，在步骤 103 确定下一个判断对象。

若在步骤 104 中判断为是，具体说在本例中字符串“Windows 操作及应用技术（第二版）”的排印信息为（FONT size=5），与周围相比为最大，因而判断为特殊排印信息。于是，进到步骤 105，在步骤 105 中将字符串“Windows 操作及应用技术（第二版）”判定为特殊字符串（商品名）。

采用本实施例的信息抽取装置，利用字号这样的排印信息来判断出特殊字符串，故能够从输入的格式文档中自动地抽取出特殊字符串。

（实施例 2）

图 3 为说明本发明的实施例 2 的文档数据和流程图。其中，图 3(a)为某个网上销售信息（HTML 形式的文档）；图 3 (b)为图 3(a)中的信息的 HTML 源文件；图 3(c)为实施例 2 的信息抽取动作的流程图。

下面说明实施例 2 的信息抽取动作的流程。与上述实施例 1 相同的动作在此省略重复的说明，仅对不同的动作进行说明。

在步骤 204 中判断在步骤 203 中确定的字符串的字体等是否与其他不同，与周围相比是否为特殊。若在步骤 204 中判断为是，具体说在本例中字符串“Windows 操作及应用技术（第二版）”的排印信息为（字体“华文行楷”，且颜色为红（color=#ff0000）），与周围相比为特殊，因而判断为特殊排印信息。于是，进到步骤 205，在步骤 205 中将字符串“Windows 操作及应用技术（第二版）”判定为特殊字符串（商品名）。

采用本实施例的信息抽取装置，利用字体和颜色这样的排印信息来判断出特殊字符串，故能够从输入的格式文档中自动地抽取出特殊字符串。

（实施例 3）

图 4 为说明本发明的实施例 3 的文档数据和流程图。其中，图 4(a)为某个网上销售信息（HTML 形式的文档）；图 4(b)为图 4(a)中的信息的 HTML 源文件；图 4(c)为实施例 3 的信息抽取动作的流程图。

下面说明实施例 3 的信息抽取动作的流程。与上述实施例 1 相同的动作在此省略重复的说明，仅对不同的动作进行说明。

在步骤 304 中判断在步骤 303 中确定的字符串的字体等是否与其他不同，与周围相比是否为特殊。若在步骤 304 中判断为是，具体说在本例中字符串“Windows 操作及应用技术（第二版）”的排印信息为（字体“华文行楷”，且为粗字（<B><FONT …</B>）），与周围相比为特殊，因而判断为特殊排印信息。于是，进到步骤 305，在步骤 305 中将字符串“Windows 操作及应用技术（第二版）”判定为特殊字符串（商品名）。

采用本实施例的信息抽取装置，利用字体和粗字这样的排印信息来判断出特殊字符串，故能够从输入的格式文档中自动地抽取出特殊字符串。

（实施例 4）

图 5 为说明本发明的实施例 4 的文档数据和流程图。其中，图 5(a)为某个网上销售信息（HTML 形式的文档）；图 5(b)为图 5(a)中的信息的 HTML 源文件；图 5(c)为实施例 4 的信息抽取动作的流程图。

下面说明实施例 4 的信息抽取动作的流程。与上述实施例 1 相同的动作在此省略重复的说明，仅对不同的动作进行说明。

在步骤 404 中判断在步骤 403 中确定的字符串的字体等是否与其他不同，与周围相比是否为特殊。若在步骤 404 中判断为是，具体说在本例中字符串“Windows 操作及应用技术（第二版）”的排印信息为（颜色为红（color=#ff0000），且为粗字），与周围相比为特殊，因而判断为特殊排印信息。于是，进到步骤 405，在步骤 405 中将字符串“Windows 操作及应用技术（第二版）”判定为特殊字符串（商品名）。

采用本实施例的信息抽取装置，利用颜色和粗字这样的排印信息来判断出特殊字符串，故能够从输入的格式文档中自动地抽取出特殊字符串。

以上的实施例 1-4 仅仅是用来说明本发明的，而不是限定本发明的。在不脱离本发明的精神实质的范围内的变更应包含在本发明中。例如，将上述实施例 1-4 进行适当组合和变更，同样可以达到本发明自动地抽取出特殊字符串的效果。

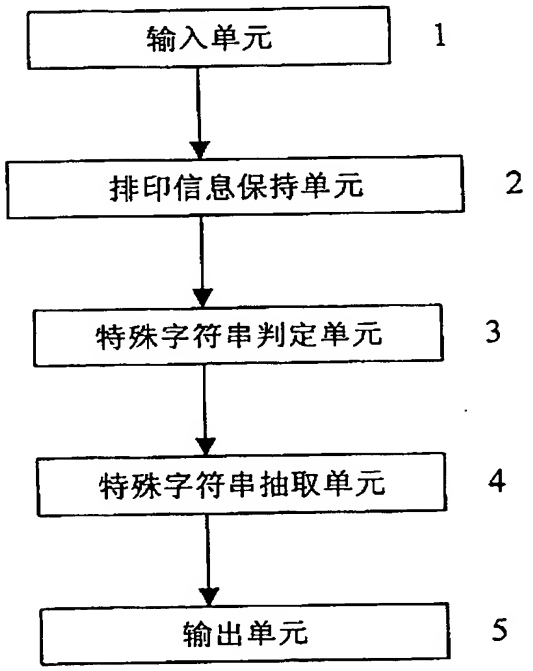

图 1

图 2

图 2(a)

图书信息

Windows操作及应用技术（第二版）

作者：肖金立
开本：16
装帧：
页数：240
出版社：电子工业出版社
出版时间：97-8-1
ISBN：7505340492

图 2(b)

```
<FONT size=5>Windows 操作及应用技术（第二版）
           
      </FONT></P>
            <P> </P>
          </TD>
        </TR>
        <TR>
  <TD class=main16px vAlign=top width="18%"><B>作者：
</B>
肖金立 <BR>
            <B>开本：</B>16 <BR>
```

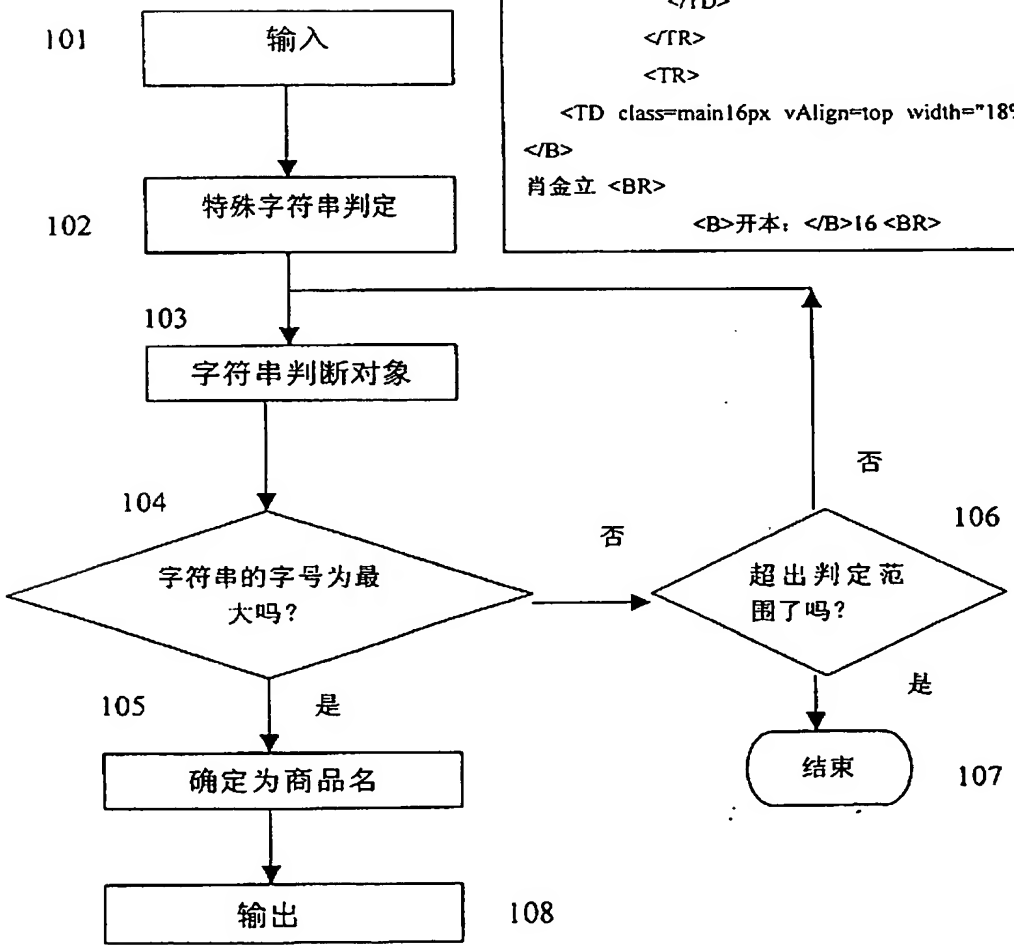

图 2（c）

图 3

图 3(a)

图书信息

Windows操作及应用技术（第二版）

作者：肖金立
开本：16
装帧：
页数：240
出版社：电子工业出版社
出版时间：97-8-1
ISBN：7505340492

图 3(b)

```
P><font face="华文行楷" color=#ff0000>Windows 操作及
应用技术（第二版）
        </font></P>
                <P> </P>
              </TD>       </TR>
    <TR>
   <TD class=main16px vAlign=top width="18%"><B>作者：
</B>肖金立 <BR>
                  <B>开本：</B>16 <BR>
                    <B>装帧：</B> <BR>
                  <B>页数：</B>240<BR>
```

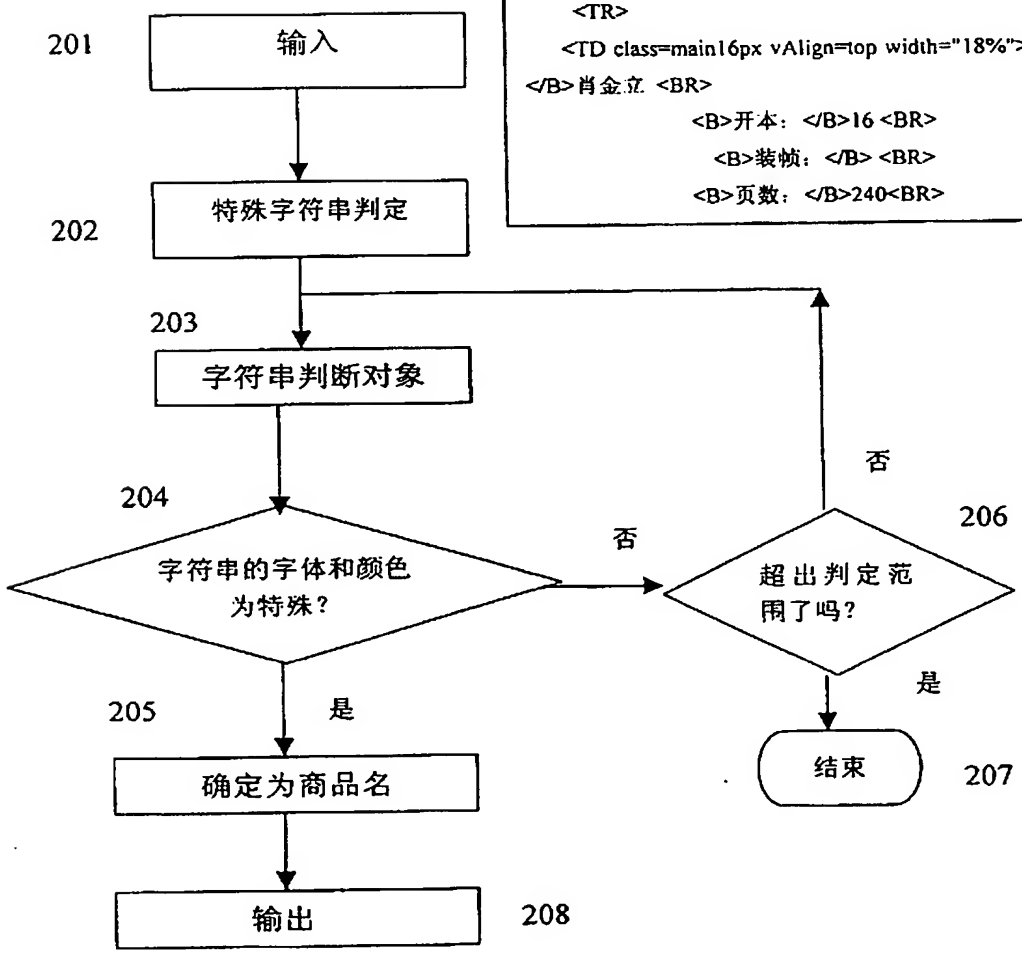

图 3(c)

图 4

图 4(a)

图书信息

Windows操作及应用技术（第二版）

作者：肖金立
开本：16
装帧：
页数：240
出版社：电子工业出版社
出版时间：97-8-1
ISBN：7505340492

图 4(b)

```
<P><B><FONT face="华文行楷">Windows 操作及
应用技术（第二版）
           
      </FONT></B></P>
            <P> </P>
          </TD>
        </TR>
        <TR>
  <TD class=main16px vAlign=top width="18%"><B>作
者：
</B>肖金立 <BR>
            <B>开本：</B>16 <BR>
```

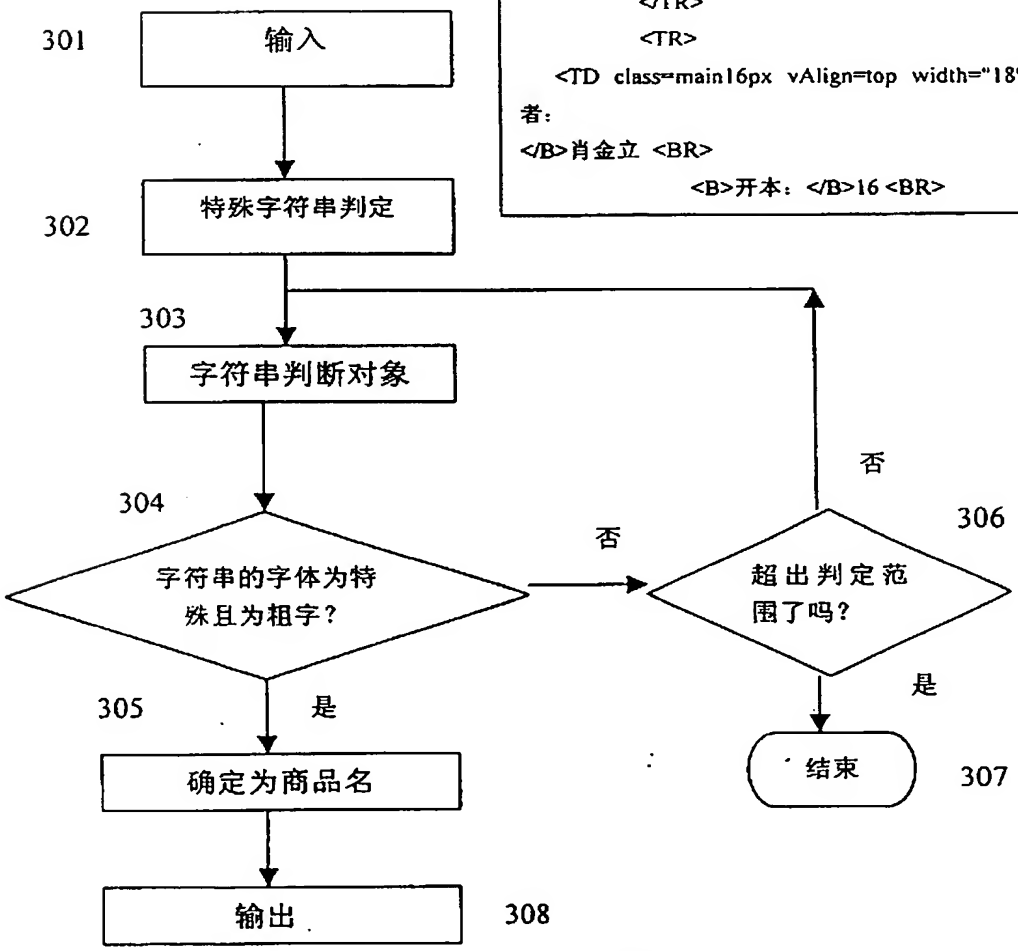

图 4(c)

图 5

图 5(a)

图书信息

Windows操作及应用技术（第二版）

作者： 肖金立
开本： 16
装帧：
页数： 240
出版社： 电子工业出版社
出版时间： 97-8-1
ISBN： 7505340492

图 5(b)

```
<P><B><FONT color=#ff0000>Windows 操作及应用
技术（第二版）
   
</FONT></B></P>
<P> </P>
</TD>
</TR>
<TR>
<TD class=main16px vAlign=top width="18%"><B>作
者：</B>肖金立 <BR>
<B>开本：</B>16 <BR>
```

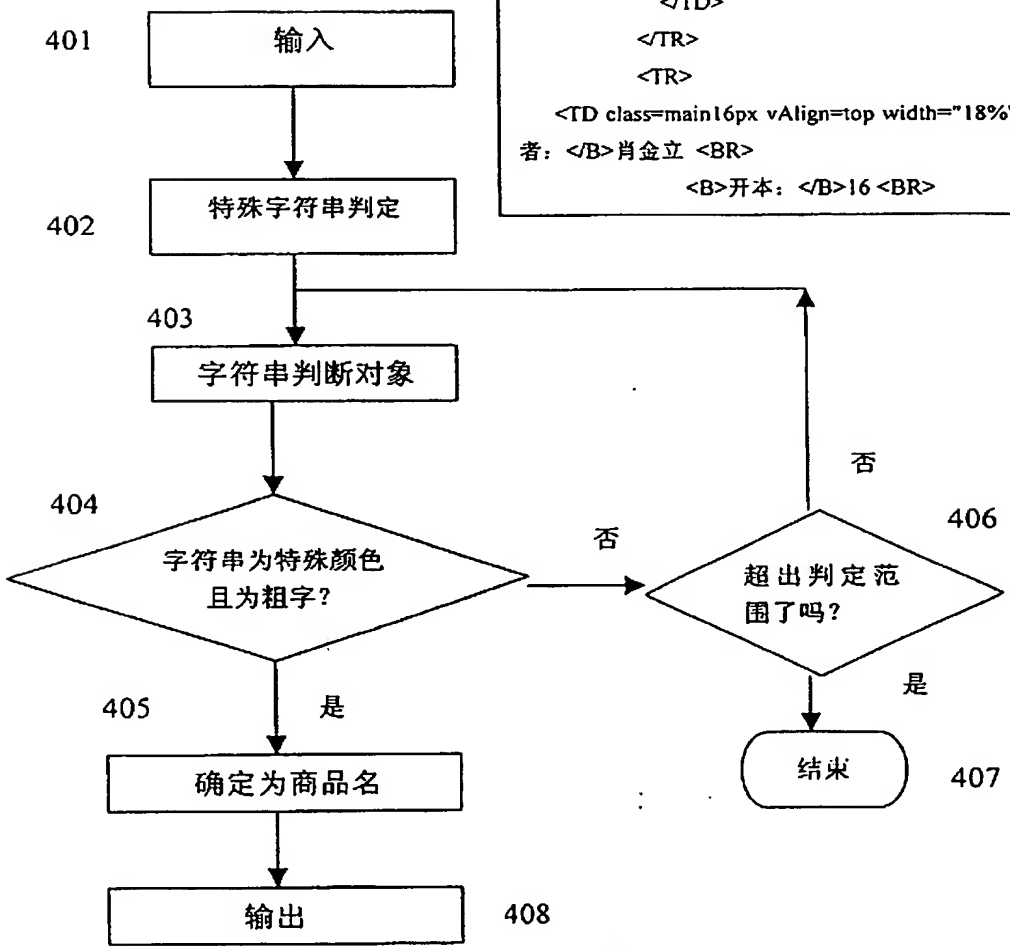

图 5(c)